DOCUMENT RESUME

ED 445 026 TM 031 631

AUTHOR Stevens, Joseph; Estrada, Susan; Parkes, Jay
TITLE Measurement Issues in the Design of State Accountability Systems.
PUB DATE 2000-04-00
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Accountability; Elementary Secondary Education; *Measurement Techniques; School Districts; *School Effectiveness; *State Programs; *Testing Programs

ABSTRACT

The practices, policies, and procedures used in all 50 states for evaluating school and school district effectiveness were examined, with a focus on the study of methodological and measurement issues in the collection, analysis, and reporting of information for accountability purposes. Data were collected through computerized literature and Web searches and by analyzing existing reports and documents. Most states are now engaged in the development of new assessment and accountability systems that can measure student, school, and district performance. Forty-eight states require some form of student testing, and many states are revamping existing testing systems and instruments, usually with an emphasis on "criterion-referenced" or "standards-based" instruments. A number of states have passed, or are considering, legislation providing monetary incentives for high-performing schools and sanctions for low-performing schools. If the current accountability movement is to be successful, more careful attention to the design and development of accountability systems will be needed. (Contains 54 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 445 026

MEASUREMENT ISSUES IN THE DESIGN OF STATE ACCOUNTABILITY SYSTEMS

Joseph Stevens, Susan Estrada, and Jay Parkes
University of New Mexico

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

J Stevens

Address correspondence to:
113 Simpson Hall
University of New Mexico
Albuquerque, NM 87131
(505)277-4203
e-mail: jstevens@unm.edu

TM031631

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April, 2000.

BEST COPY AVAILABLE

Measurement Issues in the Design of State Accountability Systems

We examined the practices, policies, and procedures used in all 50 states for evaluating school and school district effectiveness. Our primary interest was the study of methodological and measurement issues in the collection, analysis, and reporting of information for accountability purposes. Accountability occupies a central role in current educational reform efforts nationally. As described in *Quality Counts '99: Rewarding Results, Punishing Failure* (Education Week, 1999), "...accountability is the watchword, and policymakers are busy looking for ways to reward success and punish failure in an effort to improve public education" (p.3). Most states are now engaged in the development of new assessment and accountability systems that can measure student, school and district performance. Forty-eight states now require some form of student testing (Bond, Roeber, & Connealy, 1997). Many states are in the process of revamping testing systems and instruments, most commonly with an emphasis on "criterion-referenced" and "standards-based" assessment instruments. Thirty-six states now publish some form of annual report cards on individual school performance. A number of states have already passed or are considering legislation providing monetary incentives for "high-performing" schools and sanctions for "low-performing" schools.

This push for accountability reflects widespread interest on the part of the public, legislators, policy makers, and parents in obtaining information on school performance and effectiveness. There is an increasingly high-stakes environment surrounding the measurement and assessment of achievement and school effectiveness (Dorn, 1998; Heubert & Hauser, 1999). However, there are many complex issues involved in the development and implementation of these systems that require careful scrutiny and there are a number of cautions being raised about the systems being used (Elliot, 1996; FairTest, 1997; Heubert & Hauser, 1999). If the current accountability movement is to be successful, more careful attention to the design and development of accountability systems will be needed. One aspect that is critical to that development is the judicious consideration of how information is gathered, analyzed, interpreted and reported.

There are a number of challenges in attempting to measure school performance and effectiveness. The school environment is an exceedingly complex milieu in which many forces are simultaneously at work impacting the school participants. There are several kinds of "effects" on student outcomes. There are characteristics of the students themselves that have to do with their background, environment, and characteristics. There are also contextual and environmental influences on students; for example, it is clear that students with fewer resources and less opportunity have lower absolute levels of achievement. These kinds of variables must be considered in developing accountability systems. First, it is important to discover whether and how particular school policies and practices interact with these student characteristics so learning can be maximized. Second, these characteristics are factors influencing student achievement that are usually outside the control of the schooling process. As a result, schools should not be held "accountable" for their impact.

It is also worthwhile to distinguish between two kinds of "school effects". Willms and Raudenbush (1989) define Type A school effects as the total impact on a student of attending a particular school including not only what we might call quality of schooling, but also school environment, milieu, community surrounding the school, etc. Note that this definition includes all factors associated with a particular school no matter what the source. Type B school effects represent a subset of the Type A effect and include only those influences or impacts of schooling that are directly attributable to schooling practice and policy. The importance of either effect depends on one's purpose in evaluating schools. For example, for a parent, clearly the most important issue is the school with the best Type A effect. That is, the parent is interested in knowing in what school their child will achieve best *no matter what the reason*. For the purpose of monitoring the impact and effectiveness of schools, however, it should also be clear that interest should be focused on the Type B effect. That is, in schools with the same average student background and the same average school context and milieu, how effective are the practices and policies of the present school?

These distinctions are particularly important in communities where the context and composition of students and schools vary greatly. School composition and environment have been shown to have substantial effects on student outcomes over and above the effects associated with the individual student's ability and social class (Willms, 1986). For example, advantaged schools may have not only higher SES, but may show differences in parental involvement, rate of disciplinary problems, school atmosphere, peer attitudes, etc. These kinds of effects may vary from school to school and over time (Willms & Raudenbush, 1989). Thus, an acknowledgment of the complexity of the schooling process suggests that effort and care must be expended to disentangle the various determinants of student outcomes including intake characteristics of the school, school context and environment, and school policies and practice.

This suggests that an effective accountability system will need the careful choice and development of several kinds of indicators and measures that can reflect these kinds of effects on the schooling process. However, most state accountability systems have been developed rapidly; often without trial or field test, despite their high-stakes purposes. Within the last five years, most states have passed new legislation; developed new policies; adopted or developed new assessment instruments; initiated new methods for reporting educational performance; and in a number of cases, initiated systems for the reward and punishment of schools. The purpose of the present paper was to examine the methodological and measurement practices included in state accountability systems with regard to their adequacy for supporting the kinds of high-stakes inferences and decisions for which the systems are intended.

Methodological and Measurement Issues

Data on state methods, practices, and policies were collected by conducting computerized literature and web searches nationwide and by analyzing existing reports and documents that describe state practices (e.g., Bond, Roeber, & Connealy, 1997; Clements & Blank, 1997; CCSSO, 1998). These practices and policies were reviewed to determine, whenever possible, how states were using measurement principles and methods in the use and application of the

state accountability system. As a result of this review, we identified issues that occur in many state systems and that have import for the design and improvement of educational accountability systems. Discussed below are five major methodological and measurement issues: 1) Measures and indicators; 2) Design and properties of assessment systems; 3) Composites, aggregates, and indices; 4) Level and unit of analysis; and 5) Cross-sectional vs. longitudinal comparison.

Measures and Indicators. A fundamental consideration in the design of an accountability system is the choice of measures and indicators that will be gathered and used to assess performance. Fitz-Gibbon and Kochan (2000) conceptualize indicators in terms of two dimensions: 1) the kind of variable or outcomes measured (i.e., school characteristics like student numbers and school resources; quality of life information; affective outcomes; behavioral outcomes; and cognitive outcomes), and 2) the point in time when indicators are measured (intake into school; process-during schooling; immediate outcomes; long-term outcomes). While there is some variability across states in the kind of indicators and measures used, most states rely heavily on cognitive outcomes only (i.e., achievement test data) collected as a process measure during schooling. Few systems are based on an explicit design that attempts to sample from different areas of performance or from different points in time. For example, few systems use prior achievement as a control for current levels of achievement.

As states have expanded their accountability systems and developed school, district, or state report cards to disseminate information, many systems have introduced the use of additional, multiple performance indicators. In addition to achievement information, 23 states report information on student characteristics, 12 states report mobility, 16 states report on teacher qualifications, 11 report on parental involvement, 30 on student attendance and 33 states report on dropout rate. Even when additional measures are included, however, they are most commonly weighted lightly in comparison to achievement measures in the determination of levels of school effectiveness or performance.

As new indicators are incorporated into accountability systems, there is also a responsibility of the developers and users to insure that the indicators and measures are collected using common definitions and procedures across schools. Without such procedures, differences in measurement and data collection are interpreted as differences in school performance. There is also need for states to determine whether new indicators and measures are reliable, valid, and stable over time (see Mandeville & Anderson, 1987, for example). We found information and reporting on these fundamental psychometric issues to be almost entirely lacking in published state accountability information and reports. In several instances in which we have anecdotal information, some measures and indicators being applied operationally have no empirically established reliability or validity.

Design and Properties of Assessment Systems. There have been general increases in the use of several types of tests and assessments in recent years. CCSSO (1999) reports an increase in the number of states that use Norm Referenced Tests (NRT) in recent years from 29 in 1997 to 38 in 1998. There has also been an increase in the use of Criterion Referenced Tests (CRT) by the states during the same period, from 33 in 1997 to 39 in 1998. Use of writing assessments by states has declined from 39 in 1997 to 35 in 1998. In 1998, 20 states also report the use of performance assessments and 2 states report the use of portfolios. CCSSO (1999) reports that testing is pervasive across the states; all 50 states require that one or more assessments be administered to all students and 48 of the 50 states possess statewide testing systems.

Most states (44 of 50) have adopted a blended approach to assessment in their systems and use several different types of instruments and items. States report that statewide assessments are administered for three primary purposes: instructional purposes (52), school accountability (41), and student accountability (27) (CCSSO, 1999; number in parentheses indicates number reporting that purpose out of 53 states and jurisdictions).

While 48 of the 50 states require some form of statewide assessment system, almost all are in a state of flux with new assessments planned or in development. Our review of state assessment

systems shows that since 1995, major revisions of state systems is almost continual: one state has changed their assessment system 3 times, 16 states have changed systems 2 times, 26 states have changed systems once, and 3 states have maintained the same system during this four year period (the number of changes could not be determined for two other states and two additional states do not have state mandated assessments). The most common purpose of new development is to create assessments that are aligned to state content standards. Most states (42 of 48) now have some method of scoring performance against a set standard.

It is unclear, however, whether efforts to link or align assessments with content standards have been entirely successful. Most states have now adopted or are developing content standards in core academic content areas. The belief that unless assessments are aligned with state content standards they should not be used for high stakes decision making has led to substantial revision and review of state assessment systems. This has led both test publishers and test users to scrutinize existing and newly developed assessment instruments to determine their "match" to standards. While we believe this is a important process, there are a number of potential pitfalls and problems with the current practices we have observed. The determination of the alignment of an assessment instrument with standards sometimes involves simply the examination of an existing instrument and the recording of which items *appear* to be connected to particular content objectives or performance standards. We term this procedure "item-matching". Often the procedure is conducted by nonrepresentative panels of judges who may not be impartial. Unfortunately, even impartial judges may not be capable of making these appraisals accurately. A truly standards-based assessment instrument would require extensive sampling of the content domain from each relevant performance or content standard. It would also require that item development provided items at difficulty levels appropriate to the performance standard.

These are methods of test development that are common to criterion-referenced tests, and while some state assessments appear to have these properties, we are aware of a number of states in which norm referenced items and tests have been used to create standards-based

interpretations. As Popham (1998) has described, NRT tests and items are not well suited to these criterion-referenced purposes. The NRT test development process excludes many of those items that are the most relevant for criterion-referenced assessment. Thus, many of the instruments now being used in state systems that are referred to as "criterion-referenced", "standards-based", "linked", or "aligned" assessments may not in fact reliably and validly represent how well a student has mastered state content or performance standards.

In addition, in reaction to federal legislation and requirements, almost all states now have or are developing categories that translate assessment performance into a discrete number of proficiency or performance categories. Most commonly, proficiency categories are created by translating a total test score into a relatively small number of categories (usually 3-5). A number of methods have been used to set boundaries for the categories and it is well recognized that these methods are inherently judgmental (Hambleton, 1998). While standard setting methods may not be wholly objective or empirical, there are accepted methods for ensuring the quality of standard setting and there are a number of procedures that are considered to be good standard setting practice (see Hambleton, 1998). In reviewing state assessment and accountability systems, however, we have found some substantial variation in the use and application of standard setting methods to establish performance or proficiency categories. For example, while some states have used well known methods (like Angoff or Ebel), in some instances proficiency categories are set simply by taking quartiles of the distribution of a total test score on a NRT (i.e., 0-25% is well below proficient, 26-50% is proficient, etc.). In our review of state practices, we have also seen proficiency or performance categories based on scaled scores, percentile ranks, NRTs, CRTs, and performance assessments. Users of the assessments also appear to believe that if such proficiency or performance categories are reported, it automatically indicates that the assessment is "criterion-referenced" or "standards-based" regardless of the nature of the underlying assessment instrument or the methods used to define the categories.

A key concern in all of these assessment practices is whether there is sufficient evidence of reliability and validity to support the kinds of use and interpretation of information intended in the accountability system. Our review of state and national reports suggests that if such evidence has been gathered systematically, it is not apparent nor is it readily available. Even when nationally developed achievement instruments are used, information is not always complete to support certain test uses. In many cases, assessments being used are so new that there is little information on reliability and validity of the assessments. When assessments are developed locally and quickly, it is also likely that fairness issues have not been fully explored or analyzed and there may be little or no information on whether the assessments function differently for protected groups of test-takers. When proficiency and performance categories are used, it is also important that the reliability and validity of the categories is studied, not the underlying scale or total score. These are particularly important technical concerns when the assessment information is used for the high stakes purposes typical of accountability systems (Heubert & Hauser, 1999). Finally, as Messick has repeatedly argued (1989, 1994), the consequences of the assessment systems for the participants must be evaluated.

Composites, Aggregates, and Indices. Whenever multiple indicators or measurements are available, a controversy can arise as to how to use the multiple sources of information. One alternative is to create a single, global composite score or index. Often components of the index are weighted differently to reflect perceived importance of the component or curricular content area. The index is then used to judge school performance and often is also used to measure school progress, especially for Title I purposes (Adequate Yearly Progress-AYP).

A number of states now use this procedure and compute an index of school performance based on multiple measures. In our survey of state practices, we found that 32 states were using some sort of composite measure, 2 were not, and for 16 states we could not determine their current practices regarding composite indices.

If indicators and measures are highly intercorrelated, then it may be sensible to combine them into a single index of performance. When indicators and measures are relatively distinct, however, their combination may result in an uninterpretable aggregate. For example, imagine an index that is created to summarize school performance in reading, math, and attendance. It is easy to conceive of the same index score arising from a combination of low and high scores in one school and from average scores on each indicator in a second school. Some states now operationally apply indices no more sophisticated or tested than this simple example.

Depending on the scale and measurement properties of the individual measures or indicators, their combination into an index may also result in unintended weighting of the components when the measures differ in their reliability and/or standard deviations (Stevens & Aleamoni, 1986). That is, measures that differ in standard deviation become unintentionally weighted by their standard deviations when they are combined; the measure with the higher standard deviation becomes weighted more heavily. Empirical and technical study of measures and indicators should always be conducted to validate an index or composite score before it is used operationally, especially for high-stakes purposes.

Another technical concern in creating indices or composites occurs when indicators or measures are used that differ substantially in their reliability and validity. If this is the case and the measures are combined into a composite score or index, the reliability and validity of the composite may be less than the reliability of individual measures or indicators. In essence, the quality of the information provided by the best indicators may be "watered down" by poor measurement qualities in other indicators.

A related issue involves the interpretability of information used in an accountability system and the connection between indicators and standards. If performance standards are set with respect to particular measures or indicators, then it makes little sense to use an index or composite score that can not be related back to the standards. Doing so undermines the evaluation of whether the particular standards are being met. For example, a school performing well below

standard on one indicator and well above standard on a second indicator will appear to be performing on an index score at the same level as a school performing at average on both indicators. A more defensible approach in this instance is to use separate criteria for performance for each indicator. Either the use of composite indices or separate performance criteria raise a host of complex measurement and interpretation issues (Mehrens, 1990; Schmidt & Kaplan, 1971) that should be carefully addressed in the design of an accountability system to enhance validity.

Level and Unit of Analysis. Another issue of importance regards the unit of analysis in the statistical treatment of accountability data. While there has been recognition of problems in level of analysis for many years (Aitken & Longford, 1986; Burstein, 1980; Cuttance, 1992; Rowe, Hill, and Holmes-Smith, 1995) and while a number of methods are available for analysis at multiple levels (Gray, Jesson, Goldstein, Hedger, & Rasbash, 1995; Raudenbush & Bryk, 1989; Raudenbush & Willms, 1995; Scheerens, 1993), there is no recognition of this issue in most accountability systems in this country and no mechanism for treating multilevel data. Brown (1994) asserts that "...a particularly important development over the last decade has arisen from the large measure of agreement that account must be taken of differences among pupils when they arrive at school, and that the unit of analysis has to be the individual pupil rather than some average measure across pupils" (p. 56). Nonetheless in most states all analyses are conducted using school aggregates.

In accountability applications, the relevant data often occur at different levels or for different sampling "units" as in the measurement of students within schools. In this situation, there are many students each associated with a particular school. In statistical terminology, students are nested within schools. If there are relevant data for both the students and the schools, it has long been recognized that traditional analysis methods are inappropriate (Cronbach, 1976; Cronbach & Webb, 1975). If data are analyzed at the student level, the school variables are repeated exactly for each student in a school, giving a false impression of their variability. If data are analyzed at the school level (as is done in almost all state accountability systems), then all student variables

within a school must be averaged, thereby losing important information about student differences. Neither analytic approach is correct and each will result in biased interpretations of the true relationships among the variables of interest.

Recognition of the inherently multilevel nature of much educational data has provided substantial impetus to the development of new statistical methods that incorporate the essentially hierarchical nature of data in the modeling process. Due to advances in this field (see Bock, 1989, for example), several approaches are now available that allow simultaneous analysis of data from multiple levels within a hierarchical organizational structure. When analyzing school effectiveness or performance for accountability purposes, it is likely that two or three levels of the organizational structure should be included.

Another way in which multilevel structure can create problems is that the true amounts and sources of variation in schooling are never identified because they have been improperly modeled and analyzed. This is most likely to occur when student level data are aggregated and schools are treated as sampling units in the analysis. This practice precludes the examination of a number of critically important research questions. For example, with an aggregated model, it is impossible to measure the impact of practices on student learning since learning occurs at the level of the individual student; schools don't learn, students do. Previous research on multilevel modeling has also shown that data aggregated to the school level are biased (Aitken & Longford, 1986; Bryk & Raudenbush, 1987; Kennedy & Mandeville, 2000).

Another important issue in multilevel modeling involves the estimation of variability at different levels of the organizational hierarchy. For example, where does most of the variability in student achievement occur? Within classes? Within schools? Between schools? These questions ask whether the majority of student differences arise from distinctions among individual students or from differences in schooling from one school to the next. Such questions cannot be answered unless appropriate multilevel statistical models are applied. While multilevel modeling approaches have been used internationally for some time (see for example Goldstein, 1995; Rowe, et al, 1994;

1995), in one or two states in the U.S. (see for example, Sanders & Horn, 1994; Webster & Mendro, 1997), and in selected research applications (see for example, Bryk & Raudenbush, 1989; Lee & Bryk, 1989), the majority of states use one level of analysis and reporting for accountability purposes--school level aggregation of data.

Cross-sectional vs. Longitudinal Comparison. Another substantial concern arising from our survey of state practices is the common reliance on the study of different cohorts of students in cross-sectional evaluation designs as the sole mechanism to measure constructs like "student learning", "school improvement" or "school progress". In fact, in assessing school effectiveness, the common approach advocated explicitly in Title I policy and recommended by national organizations is to assess school performance by examining the difference in school aggregate performance for *different cohorts of students* over time as a measure of change. For example, in this approach the mean fourth grade achievement test scores in a school for the year 2000 cohort of students would be compared to the mean fourth grade achievement test scores for that school for the year 1999 cohort of students. A positive difference between these two different groups of students is interpreted as school improvement and a negative difference as school failure.

There is agreement in the methodological literature, however, that cross-sectional designs that study different groups of students can shed little light on learning, improvement, or other aspects of change. Nesselroade (1991) argues that in order to adequately study change one should obtain repeated measures on the same individuals. While cross-sectional designs may provide useful and important information on the level or status of performance, these designs can not effectively address questions that are inherently longitudinal. There has been a tendency to apply cross-sectional designs to longitudinal research questions for many years, however, in part perhaps, due to substantial confusion over methods for the analysis of change. Many assessment and statistical methods texts, for example, still advise readers that difference and change scores are inherently unreliable, even though such myths have been effectively dispelled (Ragosa, 1995).

We estimate that at present no more than 4-7 states have used any form of longitudinal analysis of data. Study of different cohorts as indicators of change phenomena may produce misleading results and inaccurate conclusions. Goldstein (1988), describing school effectiveness models in Britain, stated that "...It is now recognised...that 'intake' achievement is the single most important factor affecting subsequent achievement, and that the only fair way to compare schools is on the basis of how much progress pupils make during their time at school" (p.14). This recognition is not apparent, however, in accountability systems in this country.

There is a growing literature describing new and powerful methods for the analysis of change (see, for example, Meredith & Tisak, 1990; Muthen & Curran, 1997; Plewis, 1996; Willett & Sayer, 1994) and a concomitant growth in the availability of sophisticated software for longitudinal analysis. While quite rare in state accountability systems, longitudinal analyses are now in use in several states (Sanders & Horn, 1994; Webster, Mundro, & Almaguer, 1995).

To truly reflect the aspects of education that most accountability systems seek to address, there is a need to measure *student change* rather than student status. At a minimum, the prior achievement of students needs to be included in models to allow an estimate of growth. Using longitudinal models allows the conceptualization of the most relevant and direct outcome measures of school effectiveness (learning or other changes in performance or achievement). Use of prior achievement in longitudinal models also provides a crucially important degree of control over a wealth of confounding factors that complicate the evaluation of school effectiveness. SES, for example, is likely to have a stronger determinative influence on the child's status at a single point in time than on the student's rate of growth or learning (for example in our analyses, Stevens, 1999, SES indicators are the single most important predictor of student achievement status, but are *nonsignificant* in analyses of student growth). Growth is less susceptible to background, intake, and other confounding factors. As a result, schools with lower ability students are likely to fare better when growth models are used if they are effective schools. Furthermore, use of rate of growth-type outcome measures places a focus and emphasis on the most important and relevant aspects of the educational process.

Barton & Coley (1998) note that "average score trends and cohort growth tell us different things...it does appear to be important to look at *both* measures" (p. 15). They note that Maine has the highest average NAEP score and Arkansas the lowest average score *yet their gain scores are the same from 4th to 8th grade* (emphasis added). This result is truly remarkable. The two states that differ the most in NAEP scores have actually shown the same degree of improvement in student performance from 4th to 8th grade. Entirely different conclusions (and resulting policy reactions) occur depending on which data are chosen for interpretation. This result suggests that informed interpretations depend on careful interpretation of multiple sources of information.

Even in simple change models, there are two parameters of substance: initial level of performance (comparable to an intercept in a regression equation) and the rate of growth (comparable to a slope in a regression equation). Both parameters can be of great interest to educators and the two parameters may interact or influence each other. Accountability must focus on student learning, a rate of growth parameter. Reynolds & Teddlie (2000) recommend that the study of school effects should be based on longitudinal, cohort-based data on individual children. Use of longitudinal data on individual children focuses analysis on the clear and unarguable purpose of public education to effect learning, the essential outcome measure to be considered in an ideal accountability system.

Note that the common cross-sectional model used in most accountability systems only considers the equivalent of intercept information. Therefore, only average level or absolute performance is considered and change or growth is not analyzed. As a result the best predictors of absolute level of performance are variables like parental education and SES. In contrast, study of growth allows relative comparisons rather than absolute comparisons. The question is not whether the student is at a particular level of performance, but whether the student's level of performance has improved. Note that this is a crucial distinction since even an effective school may not be able to exert much influence over the absolute level of the child's functioning but can influence the growth or learning of the child (see discussion above on Type A and B school effects above).

State accountability systems not only need to incorporate multilevel modeling into their designs but should also develop methods for measuring change in individual students over time. This does not preclude the use of cross-sectional information nor does it argue that such information is invalid or unimportant. The two kinds of designs, however, provide different information on student and school performance.

Summary and Conclusions

The push for state and district accountability systems is intended to focus attention on schools that are working, schools that aren't, and provide a means to intervene in school functioning to effect educational reform. A characteristic of many accountability systems, however, is their rapid development under highly politicized conditions (Dorn, 1998) and often without concomitant technical development. This poses a threat to the success of these efforts. If methods for the assessment of school effectiveness are not carefully developed, applied, and interpreted, judgements of school effectiveness will be flawed and the resulting administrative interventions and policy decisions will be misguided. In this event, no matter what the intended consequences of the accountability system, its actual impact on the participants is likely to be negative.

We have reviewed state accountability systems now in place with particular attention to several methodological and measurement issues that are potentially problematic. Our review shows an environment in which changes in accountability systems are made rapidly and systems undergo major revisions in most states within 2-4 years. Given the importance and complexity of these systems, one has to question whether there is sufficient time in this climate to understand, develop and validate systems before they are revised and replaced.

Our review also showed that, while states are expanding the number and kind of indicators and measures they use, most state accountability systems are characterized by an over-reliance on achievement measures that are used with no correction for other factors or prior achievement.

When other measures or indicators are included, they often have no weight in the evaluation of school accountability other than their inclusion in reporting. We were also uncertain in many cases whether states had determined the reliability and validity of indicators and measures.

We found that a number of states use indices or composites of performance that appear to be untested. Use of composites or indices requires careful consideration and analysis. In many cases, indices will obscure rather than aid the accountability process. Composites and indices should be based on empirical validation and only used when measures and indicators are highly correlated. In other situations, performance on different indicators and measures should be reported separately.

Many states are developing assessment systems that are extensive and that have components that are designed to be linked to state content or performance standards. However, we found little evidence that states were involved in processes that would allow the development of assessments that could fully support these intentions and purposes. In many cases, there appears to be some mismatch between the type of assessment methods and instruments used and the accountability purpose. Many instruments also appear to lack empirical evidence for their reliability and validity.

Lastly, few state accountability systems appear to be applying appropriate analytical methods. The study of school effectiveness and performance requires the use of multilevel, longitudinal models. Very few states, however, have applied such methods to date.

References

Aitken, M. & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society (Series A)*, 149, 1-43.

Barton, P., & Coley, R. (1998). Growth *in school: Achievement gains from the fourth to the eighth grade*. Princeton, NJ: Educational Testing Service.

Bock, R.D. (1989). *Multilevel analysis of educational data*. San Diego: Academic Press.

Bond, L., Roeber, E., & Connealy, S. (1997). Trends in state student assessment programs. Washington, DC: Council of Chief State School Officers.

Brown, S. (1994). School effectiveness research and the evaluation of schools. *Evaluation and Research in Education, 8*, 55-68.

Bryk, A.S., & Raudenbush, S.W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 10(1)*, 147-58.

Bryk, A.S., & Raudenbush, S.W. (1989). Toward a more appropriate conceptualization of research on school effects: A three level hierarchical linear model. In Bock, R.D. (Ed). *Multilevel analysis of educational data*. San Diego: Academic Press.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D.C. Berliner (Ed.), *Review of research in education (vol. 8)*. Washington, DC: American Educational Research Association.

Clements, B.S., & Blank, R. (1997). *What do we know about education in the states: Educational indicators in state reports*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Council Of Chief State School Officers (1998). *State Education Accountability Reports and Indicator Reports: Status of Reports across the States 1998*. Washington, D. C.: Council of Chief State School Officers.

Council Of Chief State School Officers (1999). *Annual Survey: State Student Assessment Programs: A Summary Report Fall 1999*. Washington, D. C.: Council of Chief State School Officers.

Cronbach, L.J. (1976). Research on classrooms and schools: Formulation of questions, design, and analysis. An Occasional Paper, Stanford, CA: Stanford Evaluation Consortium.

Cronbach, L. & Webb, (1975). Between and within class effects in a reported aptitude by treatment interaction: Reanalysis of a study by G.L. Anderson. *Journal of Educational Psychology, 6,* 717-724.

Cuttance, P. (1992). Evaluating the effectiveness of schools. In D. Reynolds & P. Cuttance (Eds.), *School effectiveness research, policy and practice*. London: Cassell.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives, 6(1),* 1-34.

Elliott, J. (1996). School Effectiveness Research and its Critics: Alternative Visions of Schooling. *Cambridge Journal of Education, 26*(2), 199-224.

FairTest. (1997). Testing our children: A report card on state assessment systems. Cambridge, MA: Author.

Fitz-Gibbon & Kochan (2000). In C. Teddlie & D. Reynolds (Eds.). *The international handbook of school effectiveness research*. New York: Falmer Press.

Goldstein, H. (1988). Comparing schools. In H. Torrance (Ed.). *National assessment and testing: A research response*. London: BERA.

Goldstein, H. (1995). *Multilevel models in educational and social research: A revised edition*. London: Edward Arnold.

Gray, J., Jesson, D., Goldstein, H., Hedger, K., & Rasbash, J. (1995). A Multi-level Analysis of School Improvement: Changes in Schools' Performance over Time. *School Effectiveness and School Improvement, 6*(2), 97-114.

Guskey, T. R., & Kifer, E. W. (1990). Ranking School Districts on the Basis of Statewide Test Results: Is It Meaningful or Misleading? *Educational Measurement: Issues and Practice*(Spring), 11-16.

Hambleton, R. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In Hansche, L.N. (Ed.). *Handbook for the development of performance standards: Meeting the requirements of Title I.* Bethesda, MD: Frost Associates.

Heubert, J.P., & Hauser, R.M. (Eds.) (1999). *High Stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Kennedy & Mandeville (2000). In C. Teddlie & D. Reynolds (Eds.). *The international handbook of school effectiveness research*. New York: Falmer Press.

Lee, V. & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62,* 172-92.

Linn, R. L., & Herman, J. L. (1997). *Standards-Led Assessment: Technical and Policy Issues in Measuring School and Student Progress* (CSE Technical Report 426). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE).

Mandeville, G. K., & Anderson, L. W. (1987). The Stability of School Effectiveness Indices Across Grade Levels and Subject Areas. *Journal of Educational Measurement, 24*(3), 203-216.

Mehrens W.A. (1990). Combining evaluation data from multiple sources. In J. Millman & L. Darling-Hammond (Eds.). *The new handbook of teacher evaluation*. Newbury Park, CA: Sage Publications.

Meredith, W. & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55(1)*, 107-122.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational measurement (3rd ed.). New York: American Council on Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23(2)*, 13-23.

Millman, J. (Ed.), (1997). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Press.

Muthen, B. O., & Curran, P.J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2(4)*, 371-402.

Nesselroade (1991). Interindividual differences in intraindividual change. In L.M. Collins & J.L. Horn (Eds.). *Best methods for the analysis of change*. Washington, DC: American Psychological Association.

Plewis, I. (1996). Statistical methods for understanding cognitive growth: A review, a synthesis and an application. *British Journal of Mathematical and Statistical Psychology, 49*, 25-42.

Popham, W. J. (1998). *Your School Should Not be Evaluated by Standardized Test Scores!* AASA Online [1999, May 5].

Quality counts '99: Rewarding results, punishing failure. Education Week, Volume XVIII, Number 17, January 11, 1999.

Ragosa, D. (1995). Myths about longitudinal research. In J.M. Gottman (Ed.). *The analysis of change*. Mahwah, NJ: Erlbaum.

Raudenbush, S.W., & Bryk, A.S. (1989). Quantitative models for estimating teacher and school effectiveness. In R.D. Bock (Ed.), *Multilevel analysis of educational data*. SanDiego: Academic Press.

Raudenbush, S. W., & Willms, J. D. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307-335.

Reynolds, & Teddlie (2000). In C. Teddlie & D. Reynolds (Eds.). *The international handbook of school effectiveness research*. New York: Falmer Press.

Rowe, K.J., Hill, P.W., & Holmes-Smith, P. (1995). Methodological issues in educational performance and school effectiveness research: A discussion with worked examples. *Australian journal of Education, 39*(3), 217-248.

Sanders, W. L., & Horn, S.P. (1994). The Tennessee Value Added Assessment Model (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311.

Scheerens, J. (1993). Basic school effectiveness research: Items for a research agenda. School Effectiveness Research and School Improvement, 4(1), 17-36.

Schmidt, F.L., & Kaplan, L.B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24*, 419-434.

Stevens, J.J. (1999). *Educational accountability systems: Issues and recommendations for New Mexico*. (Technical Report), New Mexico State Department of Education.

Stevens, J.J., & Aleamoni, L.M. (1986). The role of weighting in the use of aggregate scores. *Educational and Psychological Measurement, 46*, 523-531.

Webster, W.J., & Mendro, R.L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Press.

Webster, W. J., Mendro, R. L., & Almaguer, T. O. (1994). Effectiveness Indices: A "Value Added" Approach to Measuring School Effect. *Studies in Educational Evaluation, 20*, 113-145.

Willms, (1986). Social class segregation and its relationship to pupils' examination result sin Scotland, *American Sociological Review, 51(2),* 224-41.

Willms, J.D. (1992). *Monitoring school performance: A guide for educators*. Washington, DC: The Falmer press.

Willms, J.D., & Raudenbush, S.W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability, *Journal of Educational Measurement, 26(3)*, 209-32.

Willett, J.B., & Sayer, A.G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time, *Psychological Bulletin, 116(2)*, 363-381.